PREDICTIVE MODELING PROJECT REPORT

RANIYA ARIF

PGPDSBA.O.JULY24.A

INDEX

LIST OF FIGURES

LIST OF TABLES

1 PROBLEM DEFINITION

Analysis of Driver Variables for First-Day Content Viewership on ShowTime OTT Platform

1.1 CONTEXT

OTT services have become a vital part of the entertainment ecosystem. The global OTT market, valued at $121.61 billion in 2019, is expected to reach $1,039.03 billion by 2027, driven by a 29.4% compound annual growth rate (CAGR).

ShowTime seeks to determine the key factors that impact the first-day viewership of content released on its platform. These insights will guide strategies to enhance viewer engagement and platform performance. Specifically, the analysis will focus on understanding how factors like visitor numbers, ad impressions, trailer views, and other external variables (such as major sports events) affect the viewership of new content.

1.2 OBJECTIVE

The objective of this analysis is to identify and quantify the key factors influencing the first-day viewership of content on ShowTime, an OTT service provider.

By constructing a linear regression model, this report aims to analyse the impact of variables such as user platform engagement, marketing efforts, content scheduling, and external factors like weekends and holidays.

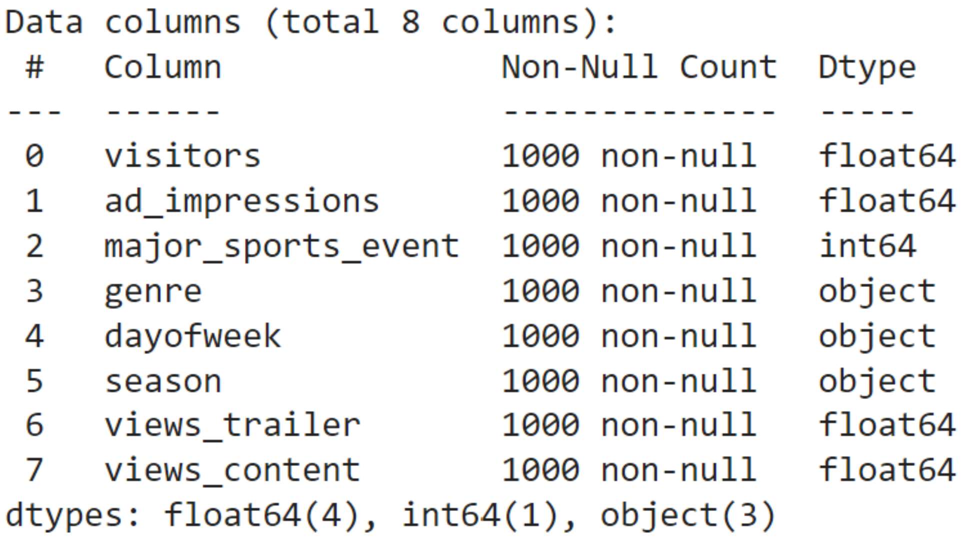
The insights from this model will help ShowTime develop targeted strategies to enhance content visibility and overall platform engagement, ensuring improved first-day viewership.

1.3 DATA OVERVIEW

The data contains the different factors to analyse for the content. The detailed data dictionary is given below.

Data Dictionary:

1. **visitors:** Average number of visitors, in millions, to the platform in the past week
2. **ad_impressions:** Number of ad impressions, in millions, across all ad campaigns for the content (running and completed)
3. **major_sports_event:** Any major sports event on the day
4. **genre:** Genre of the content
5. **dayofweek:** Day of the release of the content
6. **season:** Season of the release of the content
7. **views_trailer:** Number of views, in millions, of the content trailer
8. **views_content:** Number of first-day views, in millions, of the content



```
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   visitors            1000 non-null   float64
 1   ad_impressions      1000 non-null   float64
 2   major_sports_event  1000 non-null   int64  
 3   genre               1000 non-null   object 
 4   dayofweek           1000 non-null   object 
 5   season              1000 non-null   object 
 6   views_trailer       1000 non-null   float64
 7   views_content       1000 non-null   float64
dtypes: float64(4), int64(1), object(3)
```

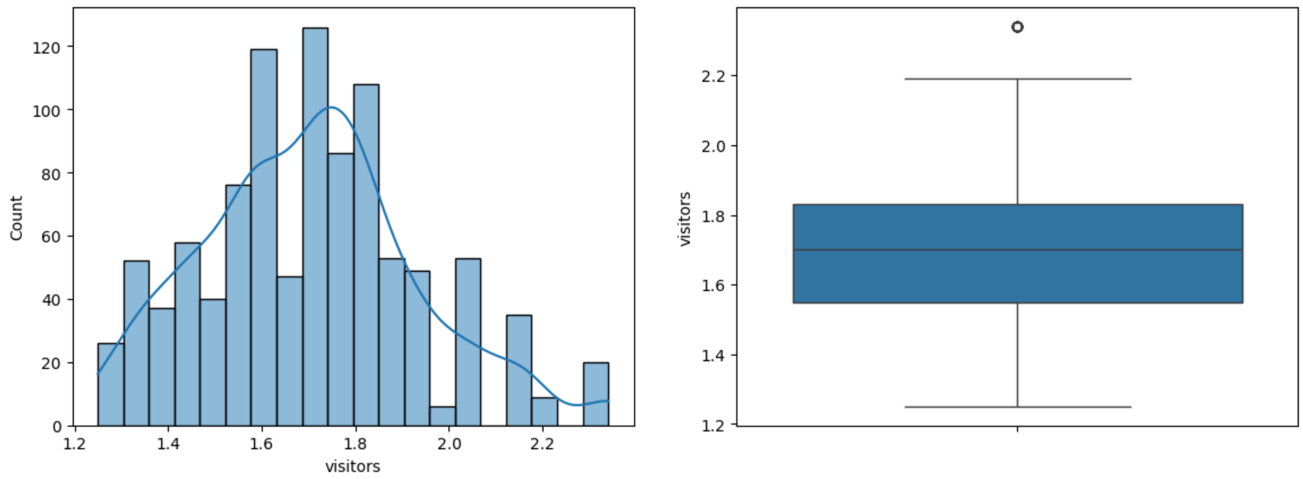
Fig. 1.3.1 Data Summary

- **Dataset Size:** The dataset comprises 1000 rows and 8 columns.
- **Data Types:**
  - **Numerical Columns:** There are 5 numerical columns in the dataset, of which 1 is of integer type and 4 is of float type.
  - **Categorical Columns:** There are 3 object-type columns, which include attributes like genre, dayofweek and season.

2. DATA ANALYSIS

2.1 UNIVARIATE ANALYSIS

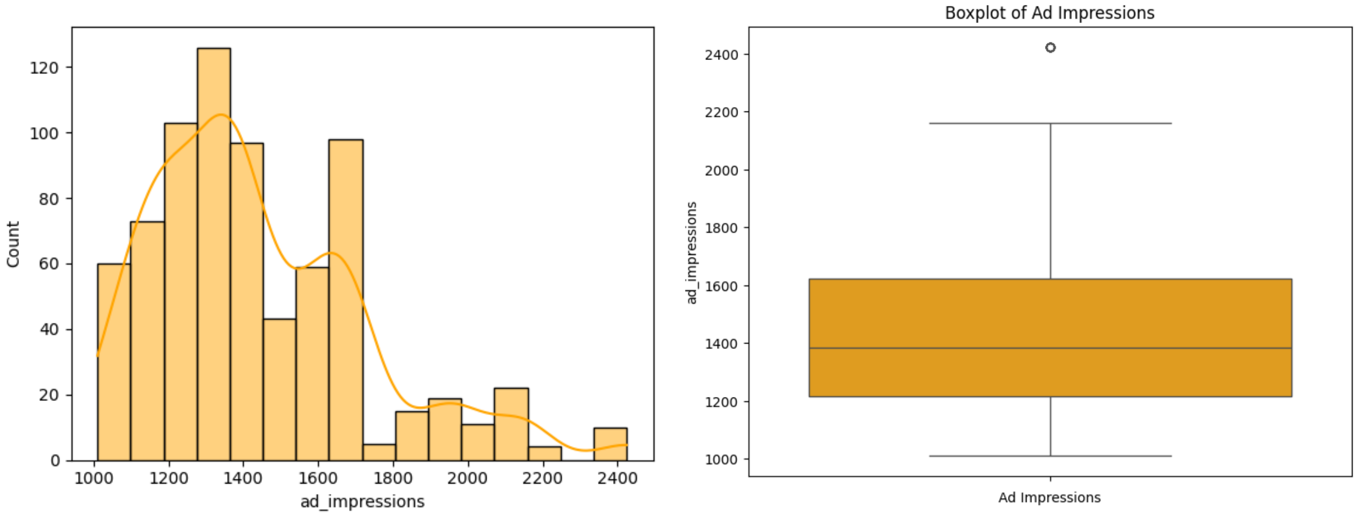
Fig. 2.1.1 Histogram and box plot of visitors



Observations :

- The histogram shows the distribution of visitors with values ranging between 1.2 to 2.2 million.
- The distribution appears to be slightly right-skewed, with a higher concentration of values around 1.5 to 1.8 million visitors.
- The majority lies between 1.6 to 1.8 million visitors, peaking around 1.7 million.
- There are fewer instances of content with either very low or very high visitor counts (i.e., below 1.3 million or above 2.0 million)
- The median visitor count appears to be around 1.7 million.
- There seem to be some outliers present on the higher end of the visitor counts, indicated by the individual data points beyond the upper whisker. This suggests that there might be a few content releases that attracted significantly higher visitor numbers than the typical content.

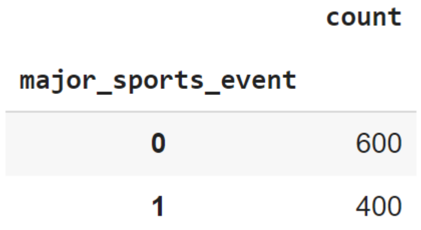
Fig. 2.1.2 Histogram and boxplot of ad_impressions



Observations :

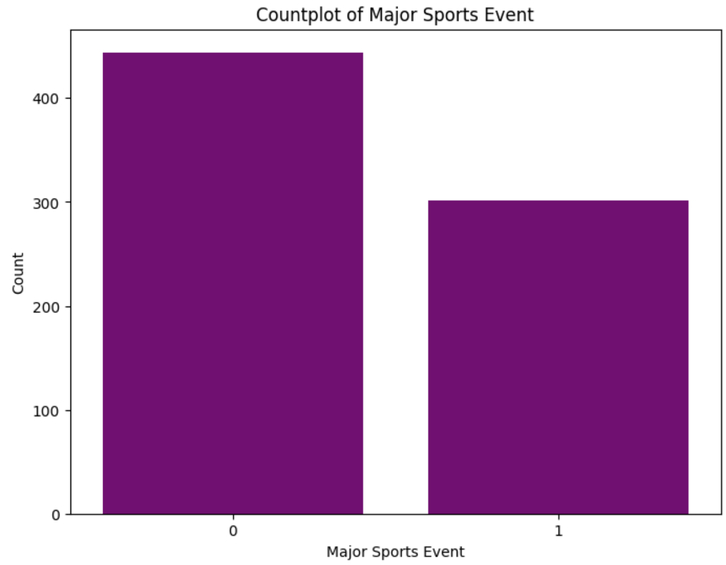
- Right-skewed distribution with most values between 1000 and 1500. Peaks around 1200 and 1400.
- Long tail extends up to 2400, indicating fewer high-impression campaigns.
- Potential outliers in the higher range, confirmed by the boxplot.

Fig. 2.1.3 Bar plot of major_sports_event



	count
major_sports_event	
0	600
1	400

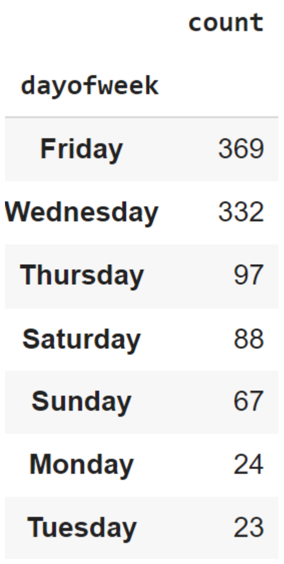
Fig. 2.1.3.1 Distribution of major_sports_event



Observations :

- The countplot shows that there are significantly more instances (600) where there was no major sports event occurring on the release date of the content.
- This suggests that the occurrence of major sports events may not be a very frequent phenomenon impacting the first-day viewership. However, it is still important to analyse the impact of major sports events on the content's viewership to understand if it significantly affects the views, especially when there is a major event scheduled.

Fig. 2.1.4 Pie chart of dayofweek



dayofweek	count
Friday	369
Wednesday	332
Thursday	97
Saturday	88
Sunday	67
Monday	24
Tuesday	23

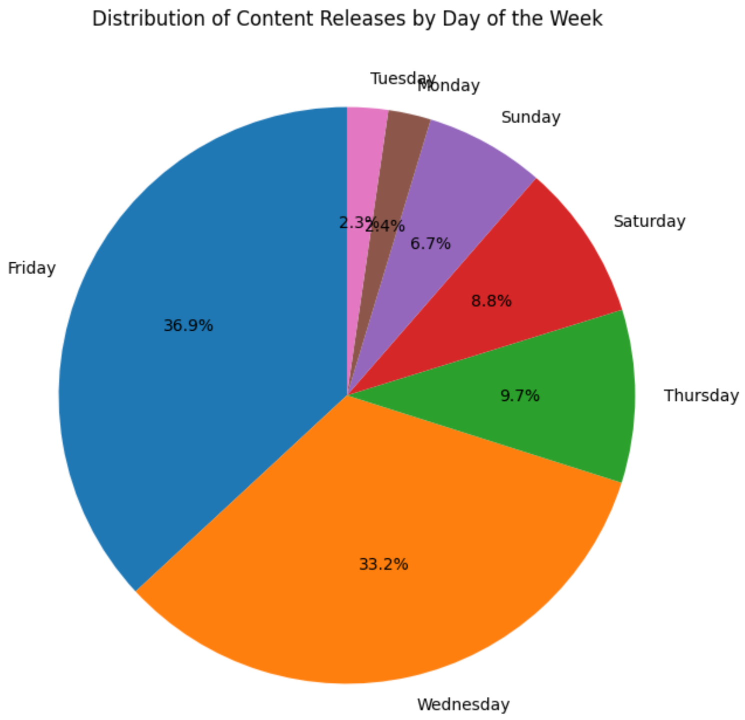
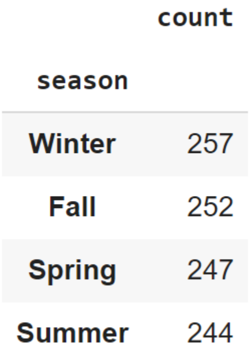


Fig. 2.1.4.1 Distribution of day_of_week

Observations :

- Majority of content releases are on Fridays followed by Wednesday.
- The least number of releases are on Mondays and Tuesdays, indicating that the viewers peak at weekends.

Fig. 2.1.5 Bar plot of seasons



	count
season	
Winter	257
Fall	252
Spring	247
Summer	244

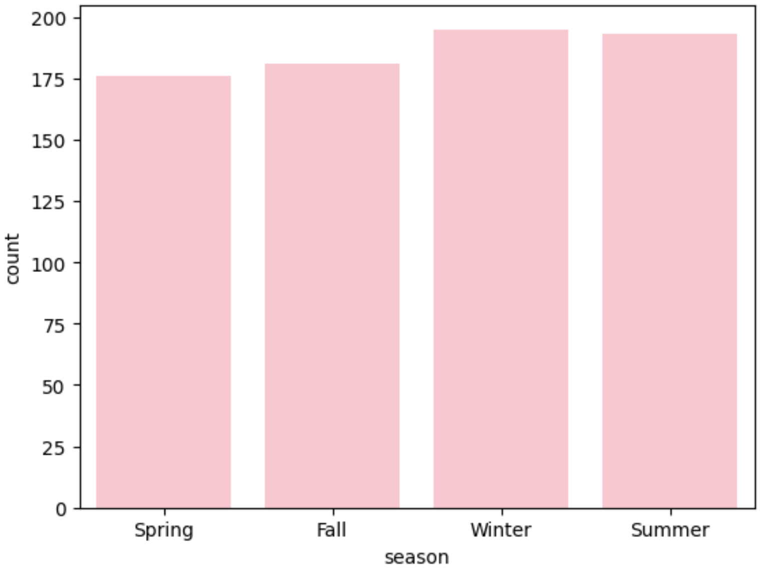
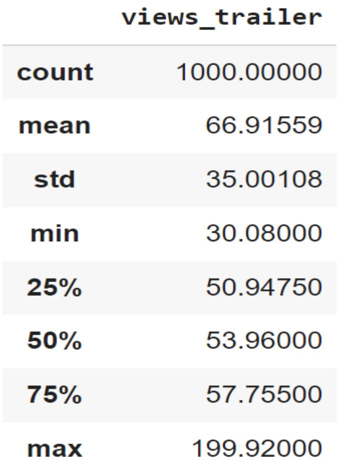


Fig. 2.1.5.1 Distribution of season

Observations :

- The content release is high in Winter and Summer seasons.
- The least content release is in Spring and Fall seasons.
- This might be because of the change in the viewer's preference for content during the seasons and holiday seasons.

Fig. 2.1.6 Histogram and Box plot of views_trailer



	views_trailer
count	1000.00000
mean	66.91559
std	35.00108
min	30.08000
25%	50.94750
50%	53.96000
75%	57.75500
max	199.92000

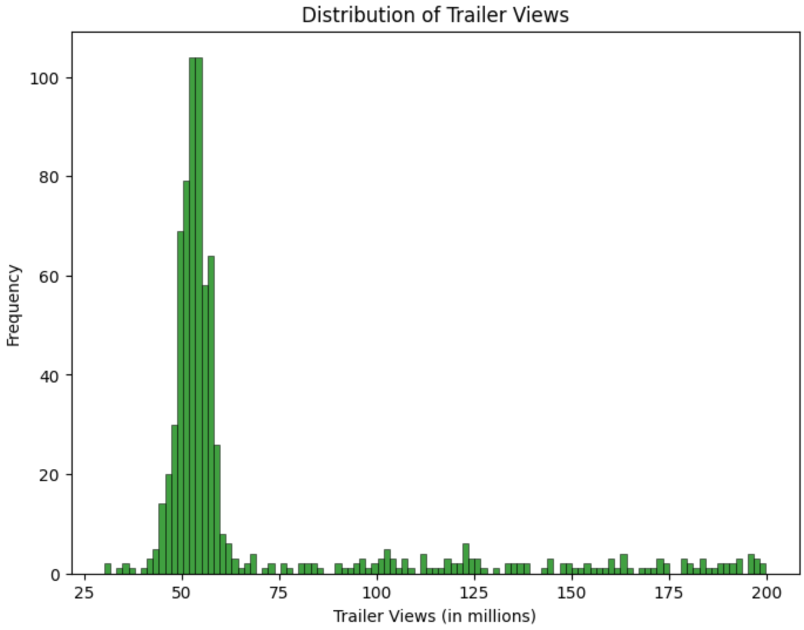
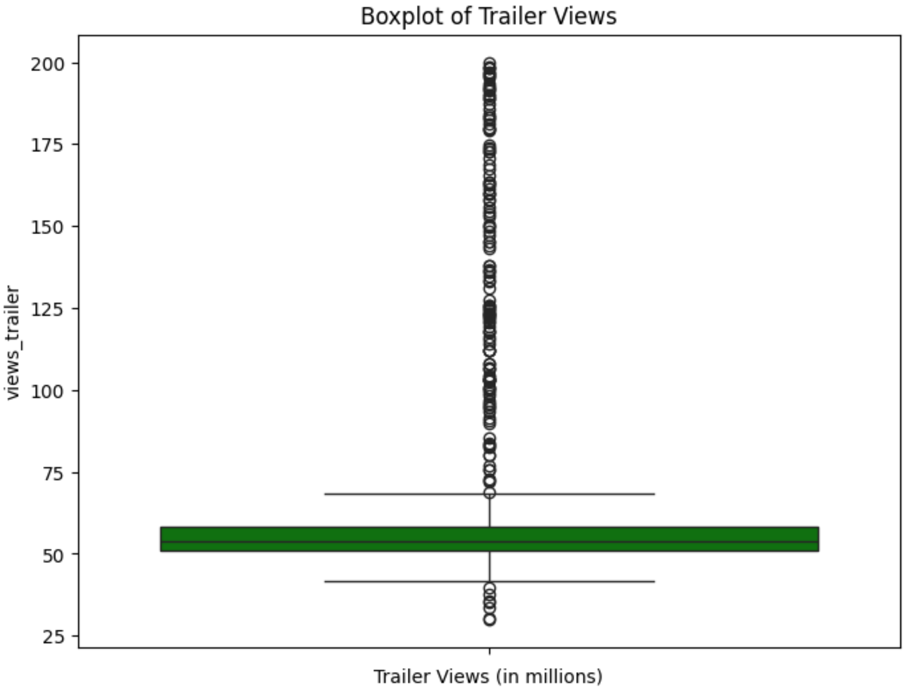


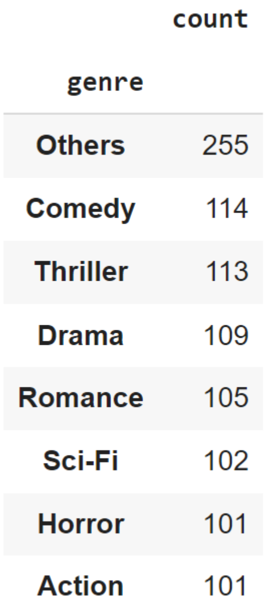
Fig. 2.1.6.1 Distribution of views_trailer



Observations :

- Central Tendency: The median trailer views are approximately 53.96 million, with a mean of 66.91 million.
- The distribution of trailer views is right-skewed, with the majority of the data concentrated between 50 million and 75 million views, indicating that a majority of content has a lower number of trailer views, while a few have significantly higher numbers.
- Outliers: The boxplot shows a significant number of outliers beyond 75 million views, with a maximum value of 199.92 million. These outliers represent content with significantly higher trailer views, potentially due to popularity or extensive marketing efforts.
- Analysing the correlation between trailer views and first-day content views can reveal whether effective trailer promotion impacts the overall viewership.

Fig. 2.1.7 Bar plot of genre



genre	count
Others	255
Comedy	114
Thriller	113
Drama	109
Romance	105
Sci-Fi	102
Horror	101
Action	101

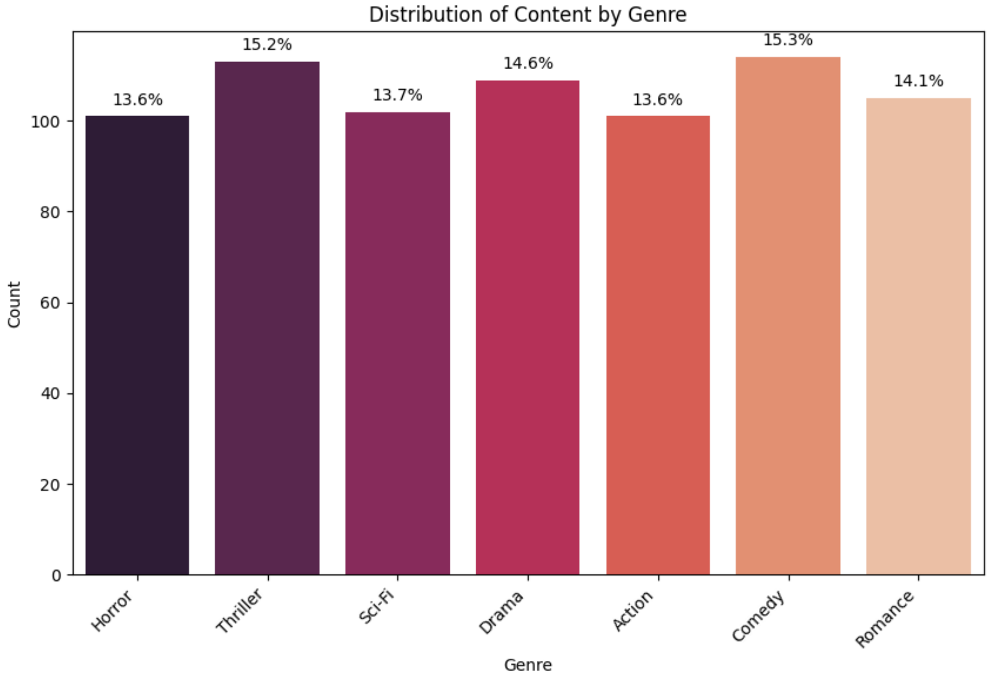
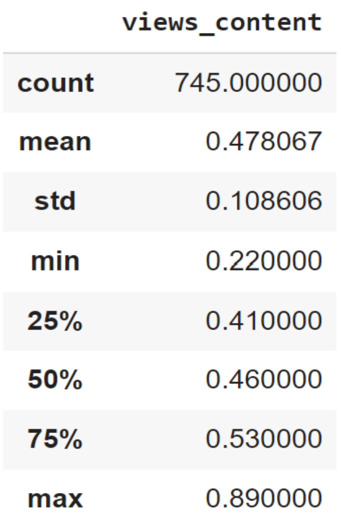


Fig. 2.1.7.1
Distribution of genre

Observations :

- Comedy appears to be the most popular genre, with the highest number of content releases, followed by Thriller and Drama
- Action, Horror, and Sci-Fi have a relatively lower number of content releases.
- Understanding the popularity of different genres can help in content strategy.

Fig. 2.1.8 Histogram and Bar plot of views_content



	views_content
count	745.000000
mean	0.478067
std	0.108606
min	0.220000
25%	0.410000
50%	0.460000
75%	0.530000
max	0.890000

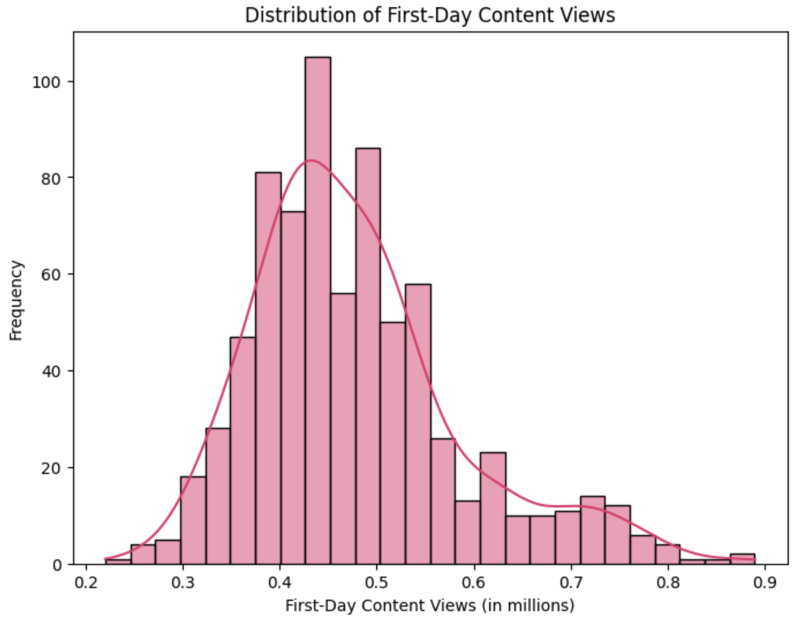
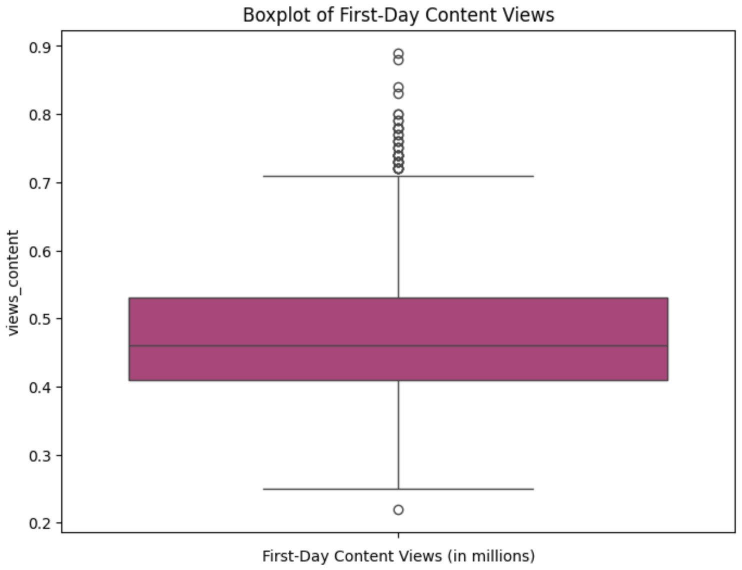


Fig. 2.1.8.1 Distribution of views_content

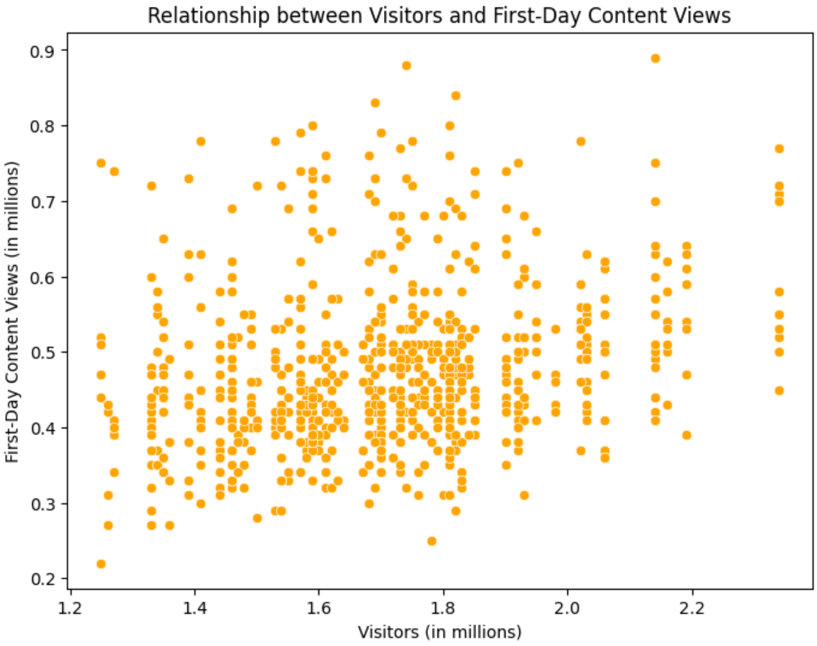


Observations :

- The distribution of first-day content views is right-skewed, indicating that a majority of content has a lower number of views, while a few have significantly higher numbers.
- The histogram shows the concentration of content views in lower ranges.
- The boxplot reveals the presence of outliers on the higher end of the distribution. These outliers represent content with significantly higher first-day views, potentially due to popularity, strong marketing, or other factors.
- Analysing the correlation between trailer views and first-day content views can reveal whether effective trailer promotion impacts the overall viewership.

2.2 BIVARIATE AND MULTIVARIATE ANALYSIS

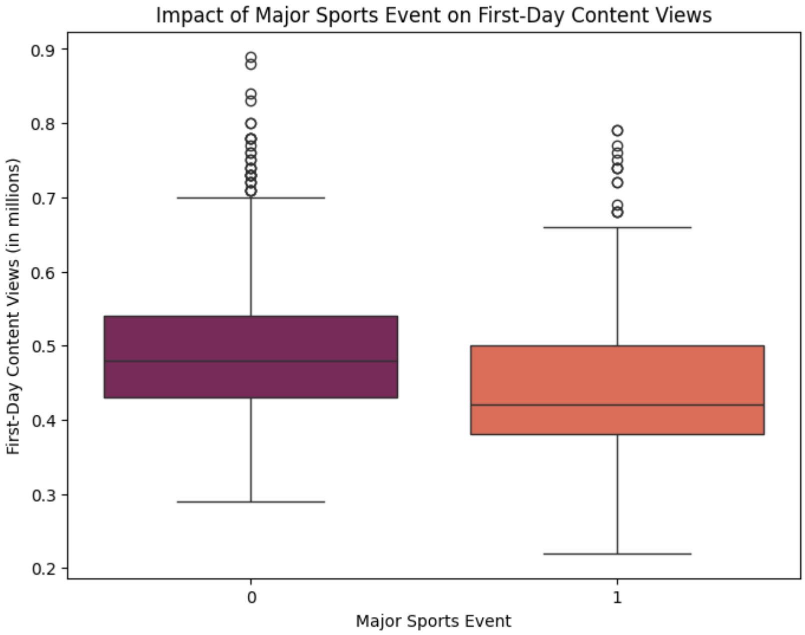
Fig. 2.2.1 visitors vs views_content

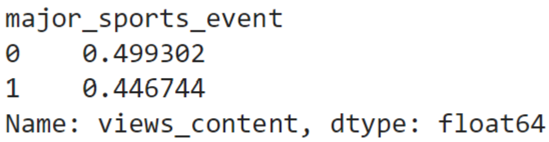


Observation :

There is no significant correlation between visitors and views_content, indicating that the number of visitors on the platform does not necessarily influence first-day viewership.

Fig. 2.2.2 major_sports_event vs views_content





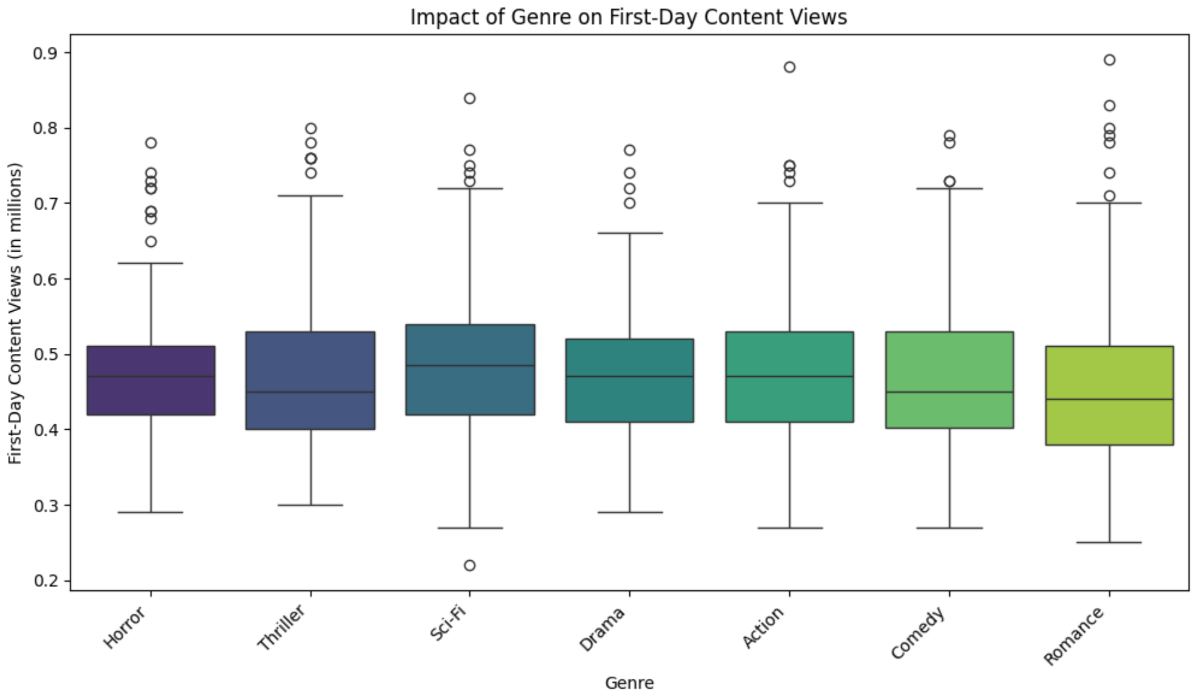
```
major_sports_event
0    0.499302
1    0.446744
Name: views_content, dtype: float64
```

Fig. 2.2.2 major_sports_event means

Observation:

- The box plot and mean 0.446744 implies that the occurrence of a major sports event has a considerable impact on the first-day content viewership.

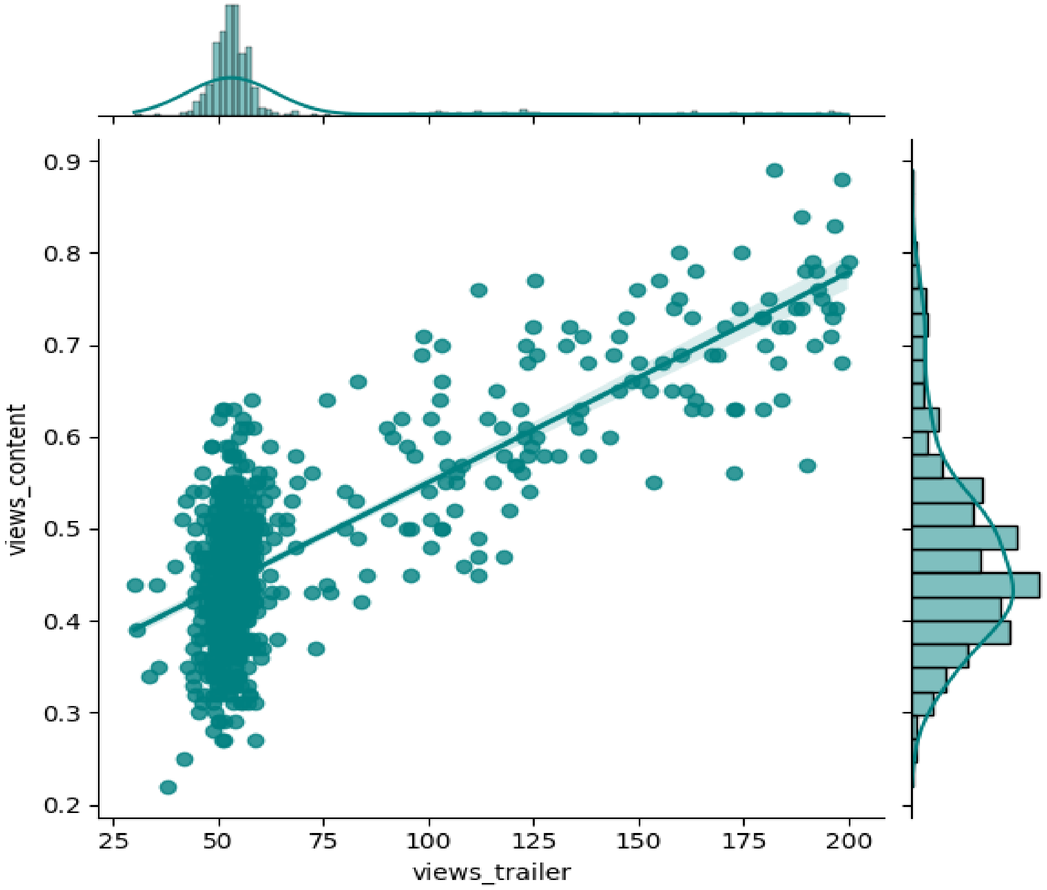
Fig. 2.2.3 genre vs views_content



Observations :

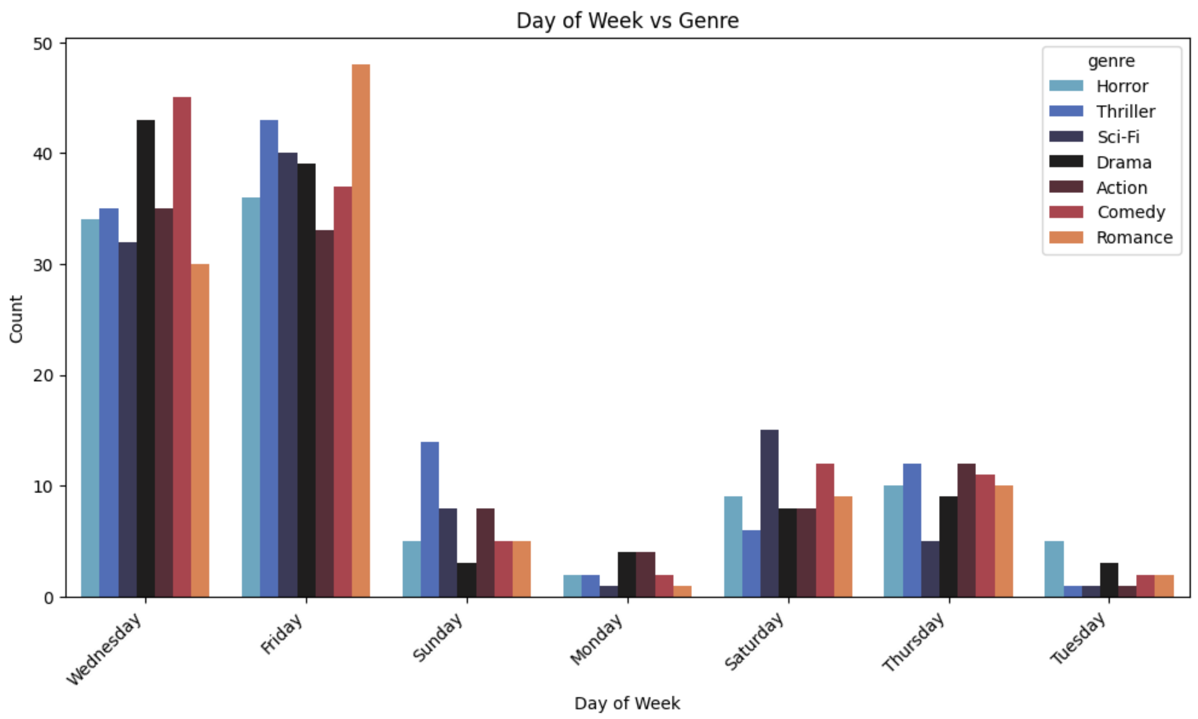
- The box plot shows that different genres have different impacts on first-day content views.
- Sci-Fi has the highest mean viewership.
- Romance has the lowest mean viewership.
- Understanding the popularity of different genres can help in content strategy.

Fig. 2.2.4 views_content vs views_trailer



- **Observations :**
- Positive Correlation: There appears to be a positive correlation between views_trailer and views_content, indicating that as trailer views increase, content views tend to increase as well.
- Dense Cluster: A large number of data points are clustered at lower values of both views_trailer (around 50 million) and views_content (around 0.4), indicating a common viewership range.
- Outliers: There are some scattered points with much higher values for views_trailer, suggesting that a few trailers had significantly more views than the majority.
- Linear Fit: The regression line indicates a moderately strong linear relationship, but there is still some spread of data points around the line, showing variability in how trailer views translate to content views.
- KDE and Histogram: The marginal distributions (histogram and KDE) show the distribution of trailer views and content views. Both seem to have a right-skewed distribution, indicating that most trailers have fewer views, with a few having significantly higher views.

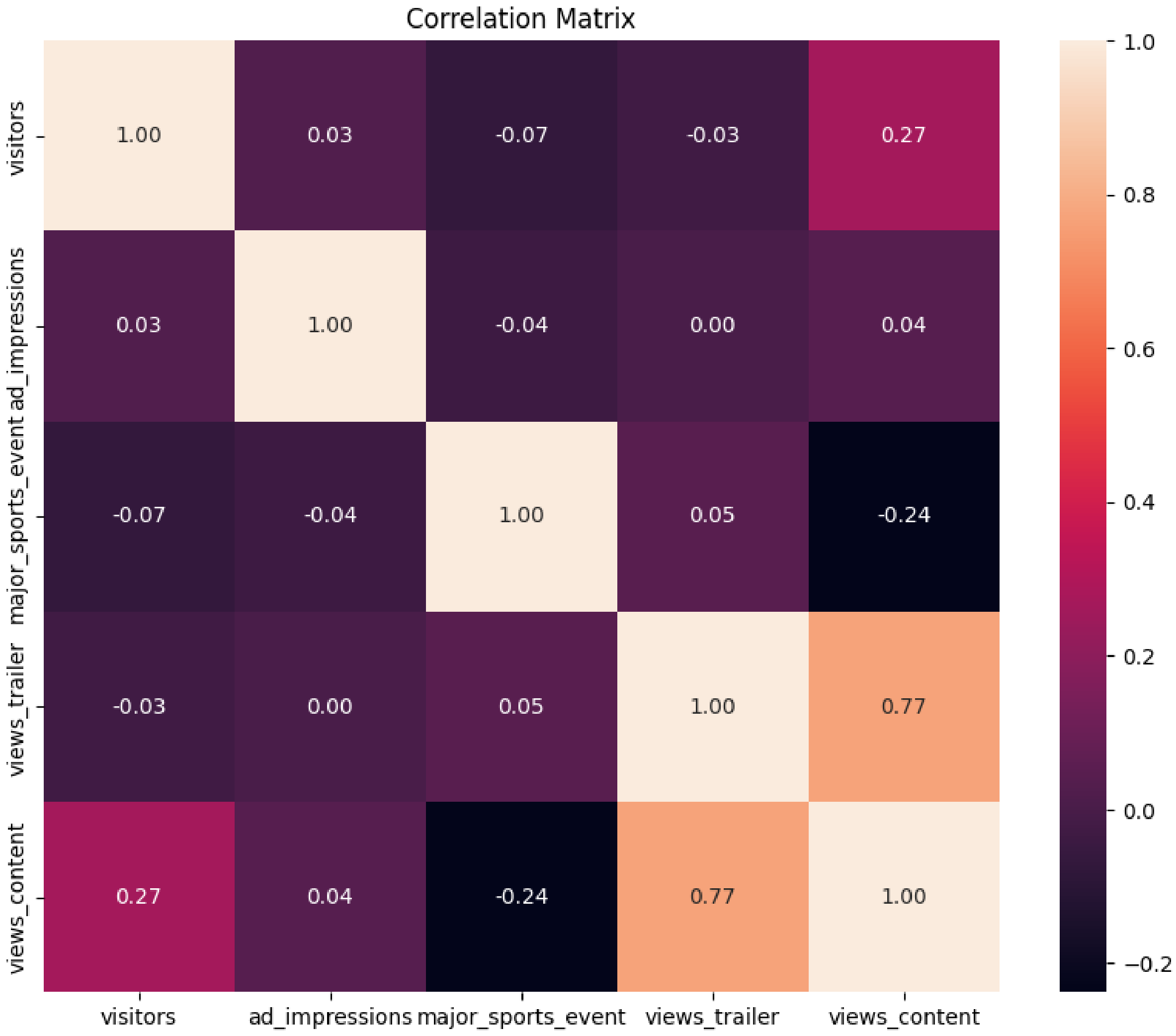
Fig. 2.2.5 dayofweek vs genre



Observations:

- Certain genres appear to be more popular on specific days of the week. Romance is the most popular genre on Fridays where as Horror is most popular on Tuesdays.
- The content scheduling strategy can be explored to align with the popularity of different genres on different days.
- Analysing patterns between genre and day of the week can help optimize content releases for maximum viewership.

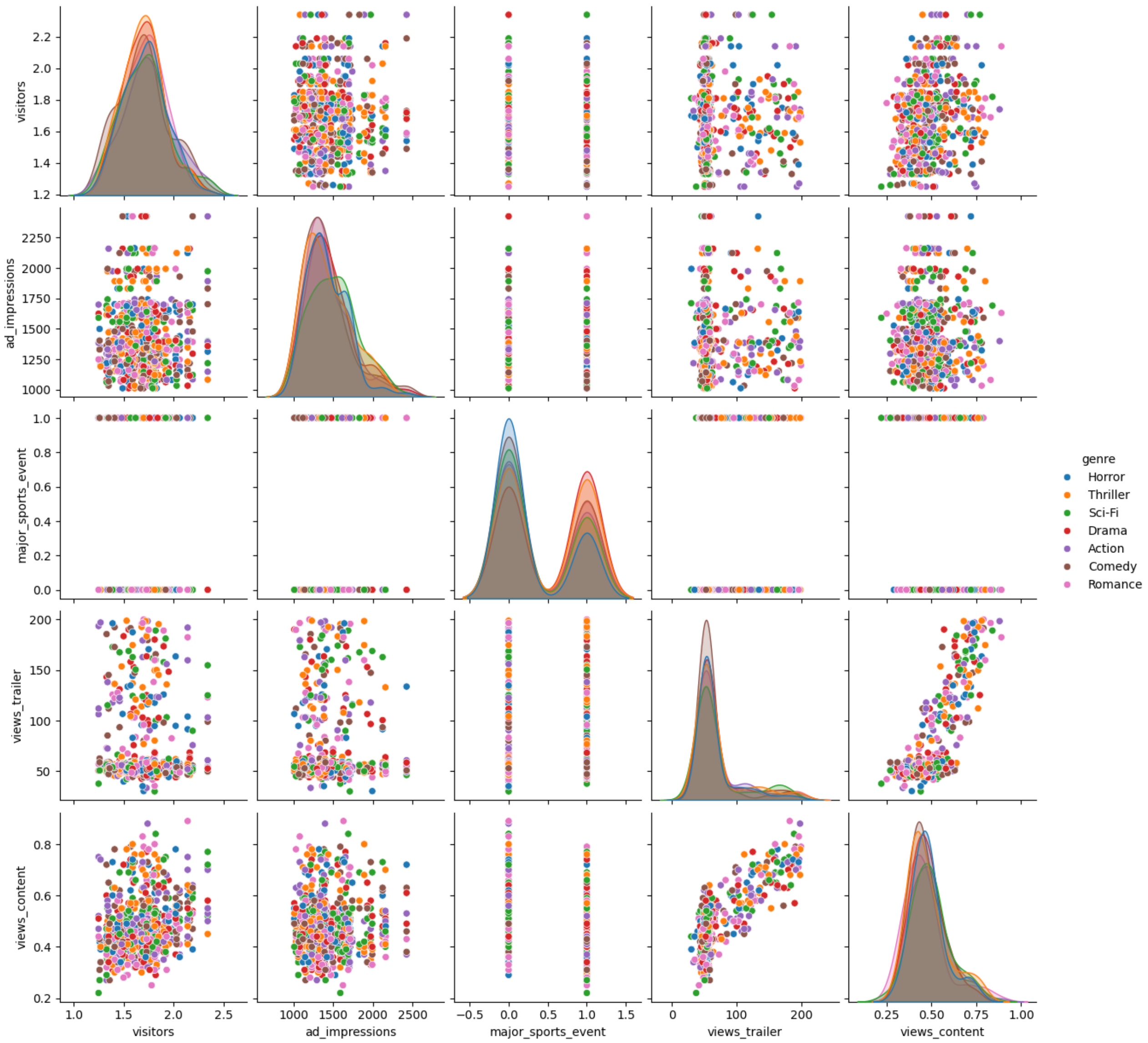
Fig. 2.2.6 Correlation Matrix



Observations:

- views_trailer and views_content have a strong positive correlation, suggesting that more trailer views lead to more content views.
- ad_impressions and views_content also show a moderate positive correlation.
- There is no significant correlation between visitors and views_content, indicating that the number of visitors on the platform does not necessarily influence first-day viewership.

Fig. 2.2.7 Pair Plot for Numerical Data By genre



3 ANSWERING KEY QUESTIONS

1. **The day of the week on which content is released generally plays a key role in the viewership. How does the viewership vary with the day of release?**

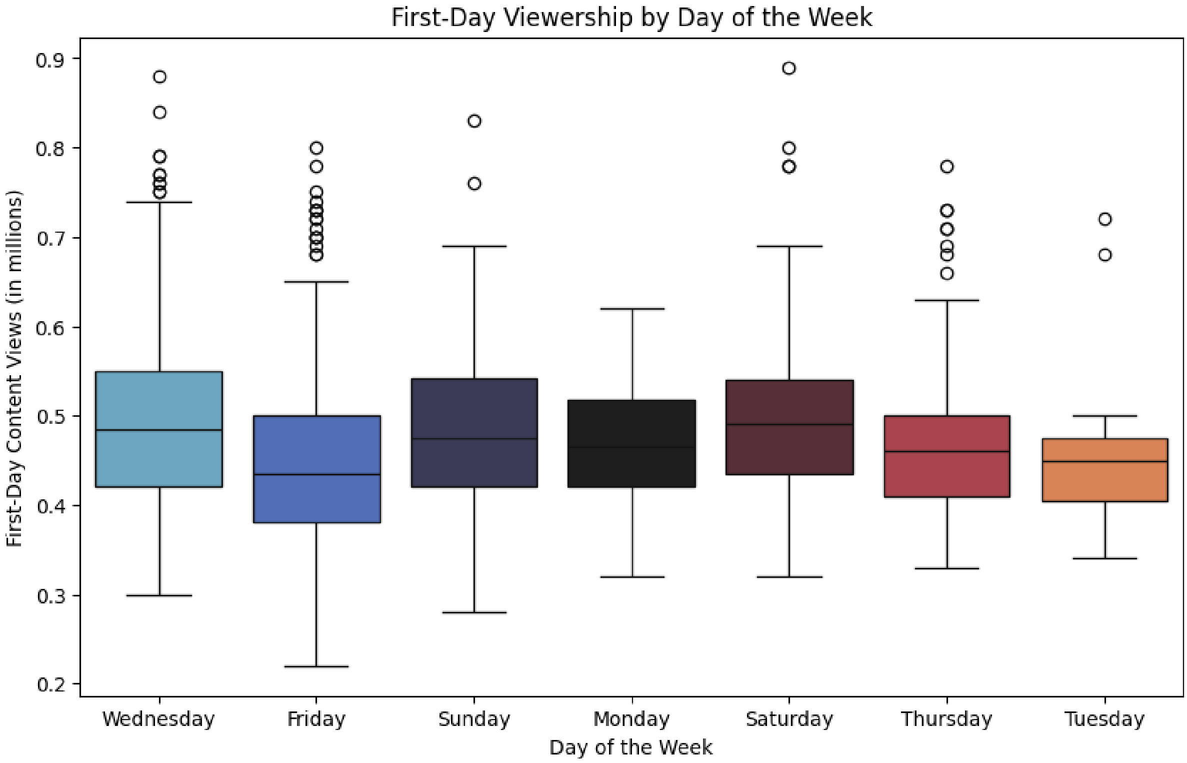


Fig. 3.1.1 Box Plot - First Day Viewership by Day of the Week

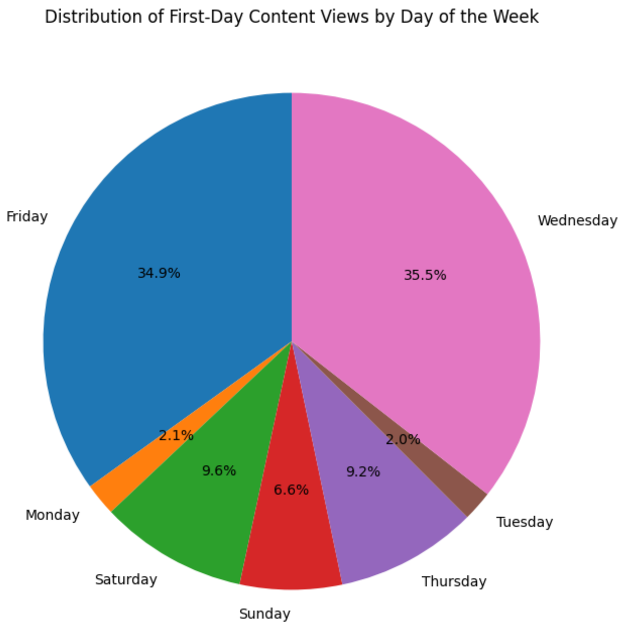


Fig. 3.1.2 Pie Chart - First Day Viewership by Day of the Week

Inferences:

- The boxplot reveals that the median first-day viewership is generally higher on weekends (Saturday, and Sunday) compared to weekdays. This suggests that viewers are more likely to watch content on ShowTime during weekends, potentially due to increased leisure time.
- Content released on weekdays like Monday and Tuesday tends to have lower viewership on the first day, as indicated by the lower median and quartiles.
- There are also more outliers on the weekends, suggesting that there are certain instances of content released on weekends that have unusually high first-day viewership.

2. How does the viewership vary with the season of release?

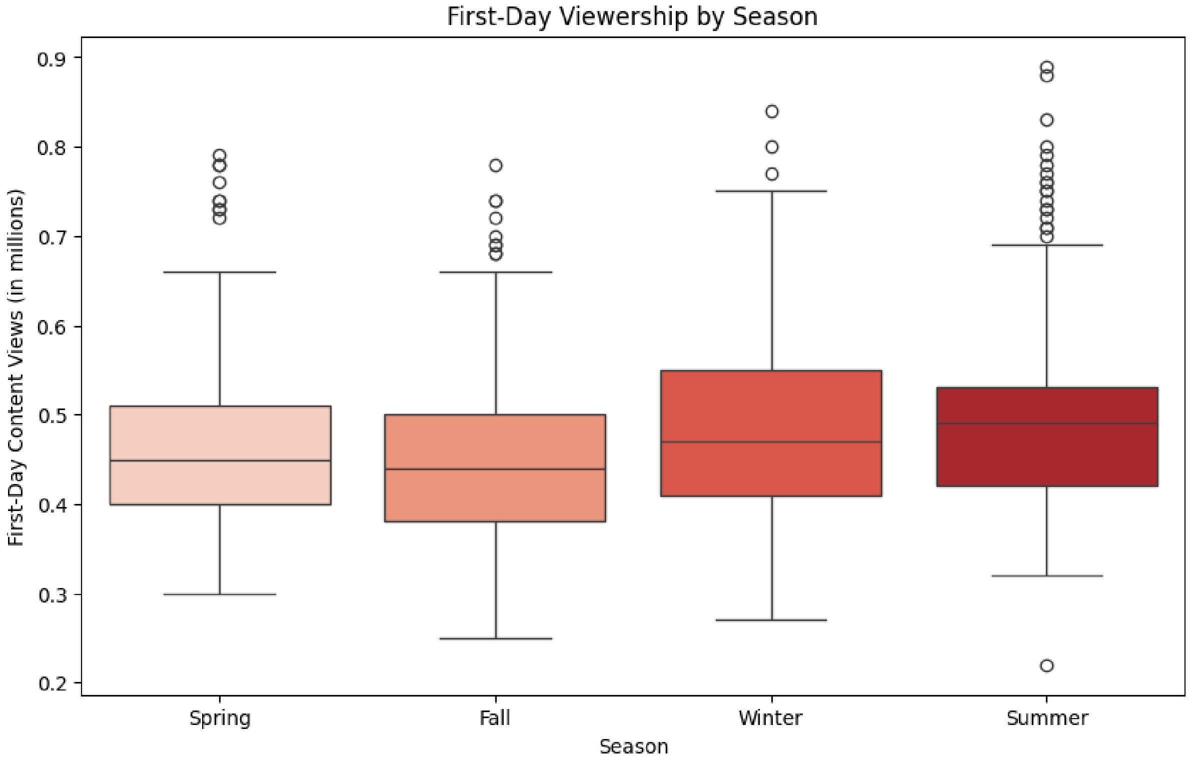


Fig. 3.2.1 Box Plot - First Day Viewership Season

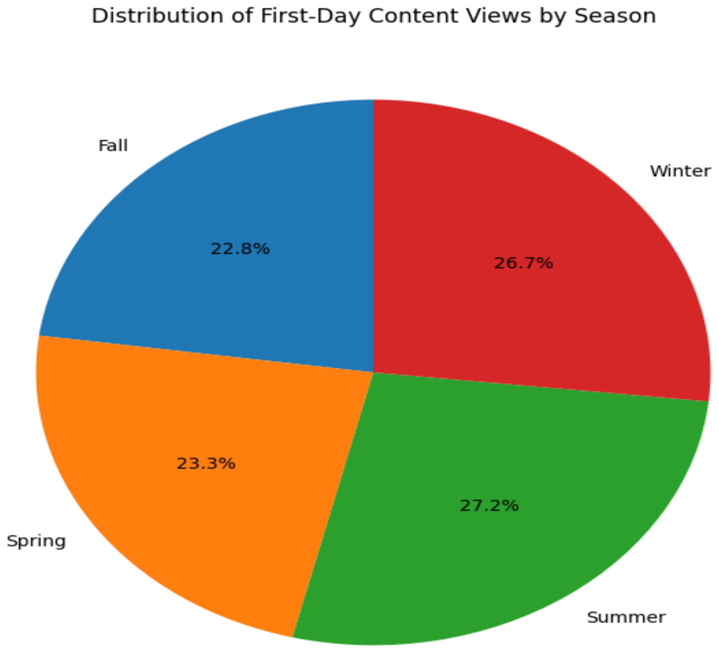


Fig. 3.2.2 Pie Chart - First Day Viewership Season

Inferences:

- Summer and Winter seasons tend to have a higher median first-day viewership compared to Fall and Spring
- This suggests that viewers might be more inclined to watch content on ShowTime during these seasons, potentially due to factors like weather or seasonal preferences
- There's also a noticeable difference in the distribution of viewership across different seasons, with Winter having a slightly wider range, suggesting more variability
- Further analysis could delve into the content type released during each season and user behaviour in different seasons to understand this pattern better

3. What is the correlation between trailer views and content views?

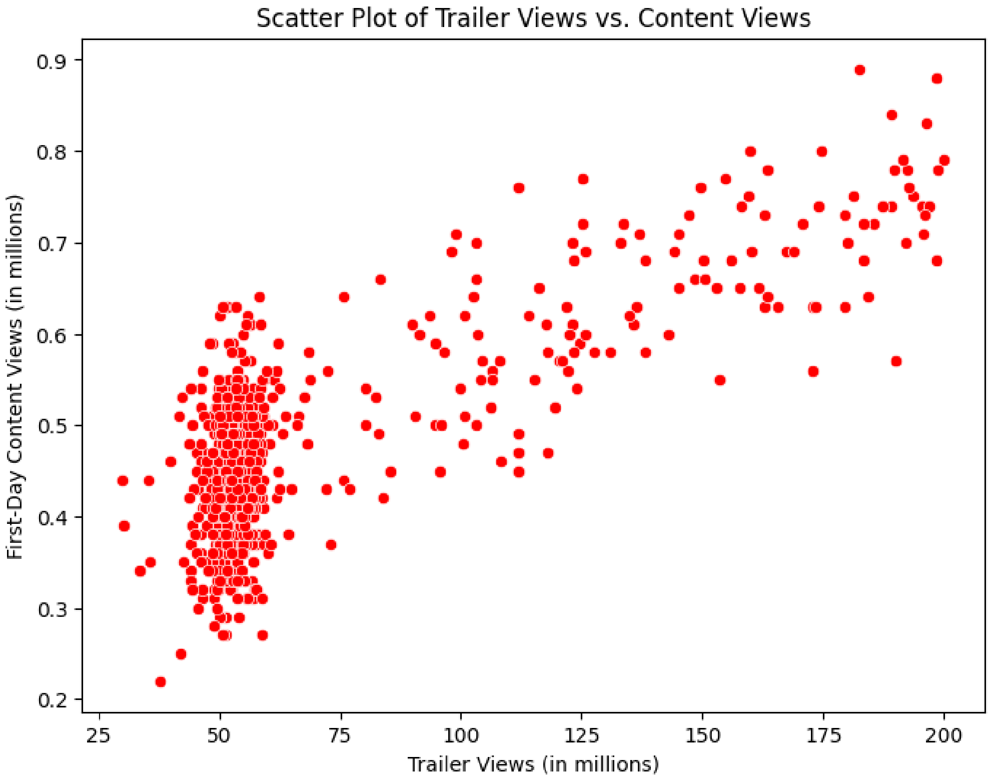


Fig. 3.3 Scatter plot of views_content and views_trailer

Inferences:

- The Scatter plot suggests there is a strong positive correlation between trailer views and content views
- This suggests that higher trailer viewership is associated with higher first-day content viewership
- Promoting content effectively through trailers can potentially drive increased viewership.

4. What does the distribution of content views look like?

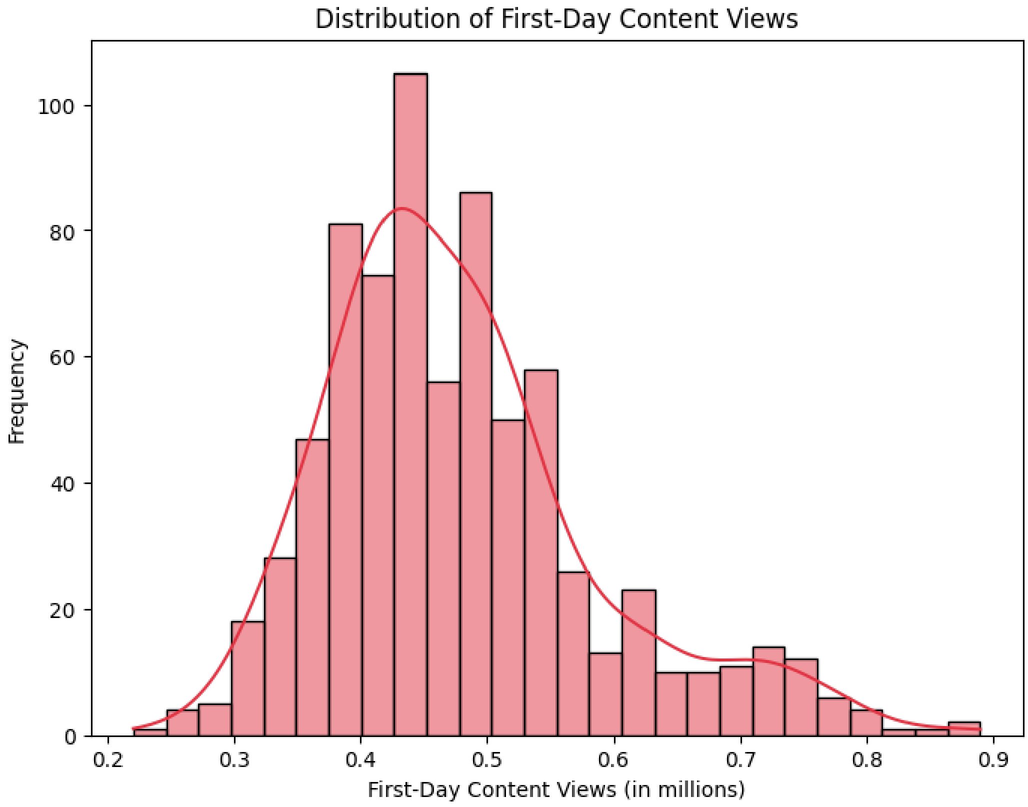


Fig. 3.4 Distribution of Content Views

Inferences:

- The distribution of first-day content views is right-skewed, indicating that a majority of content has a lower number of views, while a few have significantly higher numbers.
- The histogram shows the concentration of content views in lower ranges.
- The boxplot reveals the presence of outliers on the higher end of the distribution. These outliers represent content with significantly higher first-day views, potentially due to popularity, strong marketing, or other factors.
- Analysing the correlation between trailer views and first-day content views can reveal whether effective trailer promotion impacts the overall viewership.

5. What does the distribution of genres look like?

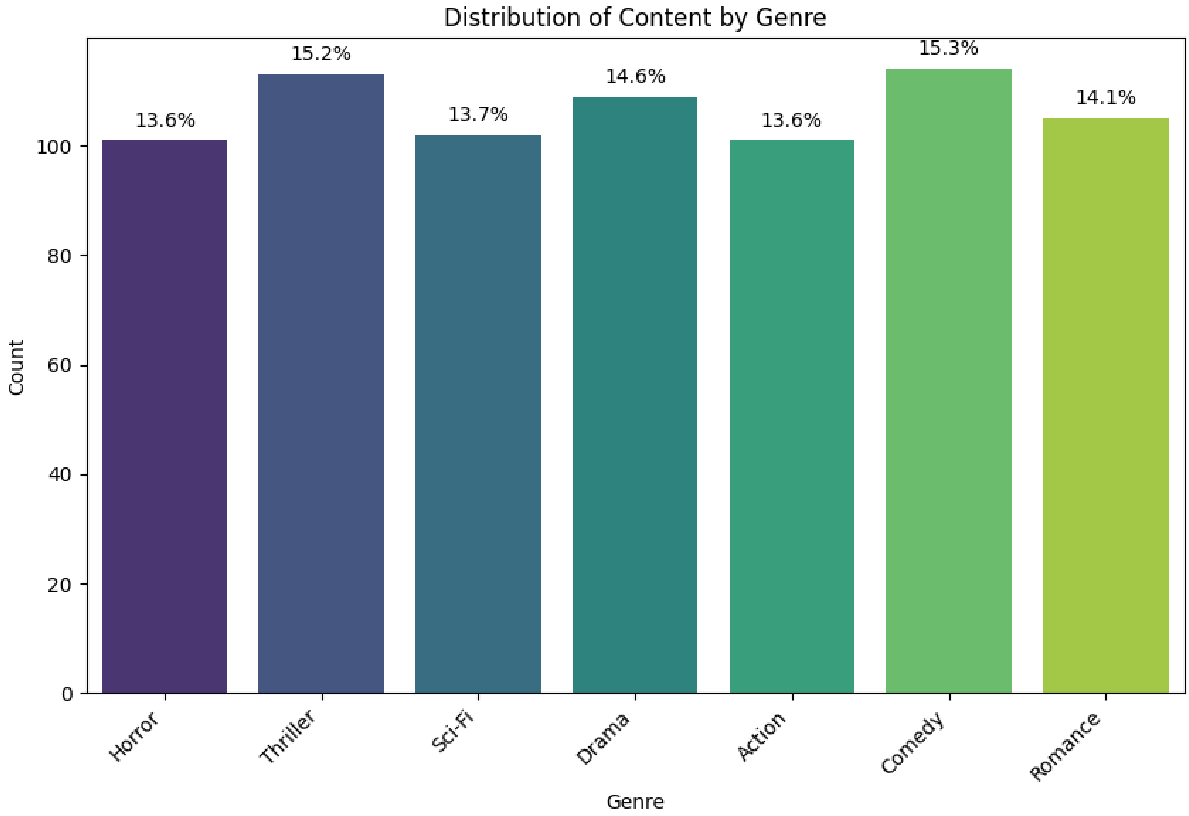


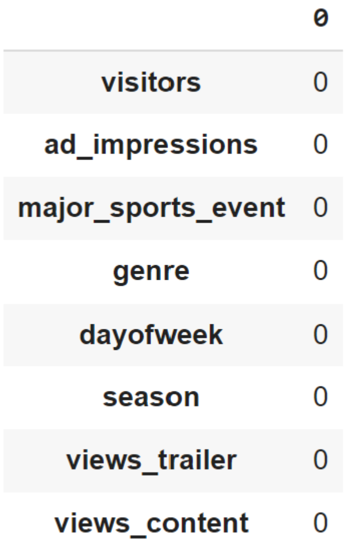
Fig. 3.5 Distribution of Genre

Inferences:

- Comedy is the most common genre, making up 15.3% of the total content.
- Thriller follows closely with 15.2%.
- Drama and Romance are also popular, accounting for 14.6% and 14.1%, respectively.
- Sci-Fi is at 13.7%, making it another frequent genre.
- Horror and Action are the least common, both at 13.6%.

4 DATA PREPROCESSING

- **Missing values:** No missing values were present across any of the columns. This ensured that no Data Imputation or Removal process were required, thereby preserving the integrity of the original data.

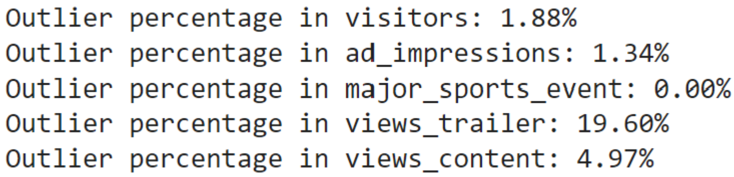


	0
visitors	0
ad_impressions	0
major_sports_event	0
genre	0
dayofweek	0
season	0
views_trailer	0
views_content	0

Fig. 4.1 Distribution of Missing values

- **Duplicate values**: Upon thorough examination of the dataset, no duplicate values were identified. This indicates that all records are unique, ensuring data integrity and reliability for subsequent analysis.

- **Outliers Detection and Treatment**



```
Outlier percentage in visitors: 1.88%
Outlier percentage in ad_impressions: 1.34%
Outlier percentage in major_sports_event: 0.00%
Outlier percentage in views_trailer: 19.60%
Outlier percentage in views_content: 4.97%
```

Fig. 4.2 Outlier percentages in each column

Outliers represent extreme cases of high viewership, which could potentially skew the model's interpretation of what affects viewership on average. Since the goal is to understand the general factors influencing first-day content views, capping outliers can help focus on more typical cases, allowing clearer insights into factors like major sports events, weekdays, and seasons.

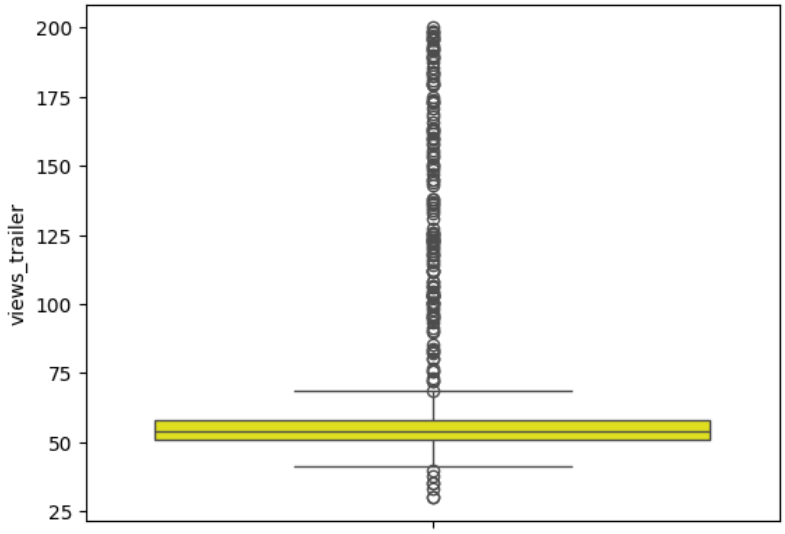


Fig. 4.3 Outliers in views_trailer column

Since the dataset contained only 1000 rows and outliers were 19.6% in the views_trailer column, direct removal would lead to loss of significant amount of data. Therefore, the extreme values were replaced with lower and upper bounds generated through the IQR method.

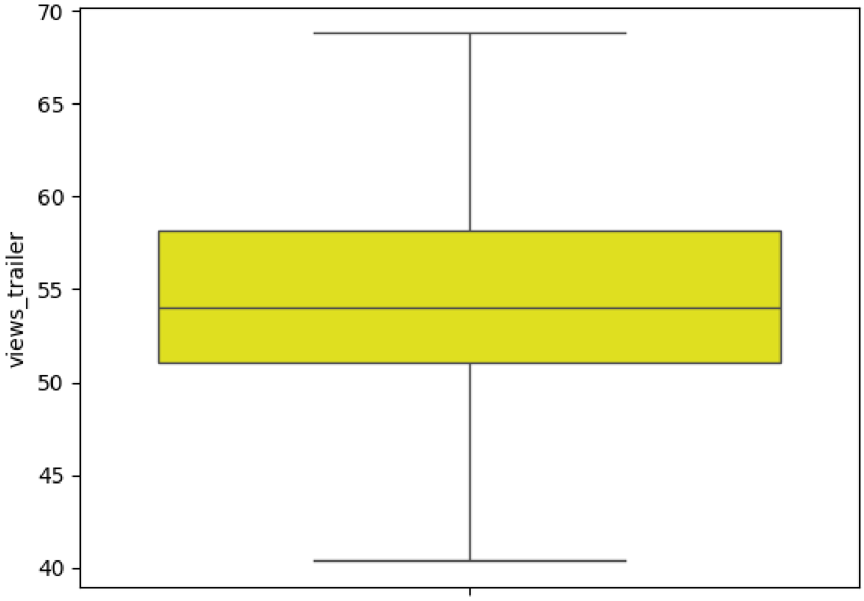
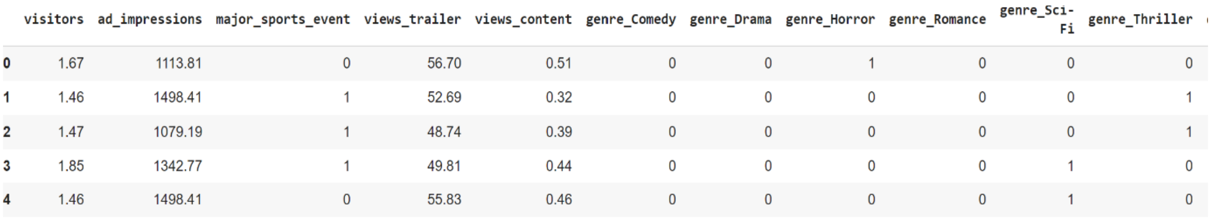


Fig. 4.3.1 Box plot of views_trailer after removing outliers

Similarly, the outliers in views_content column were handled.

- **Data Transformation**

The categorical columns, ‘dayofweek’, ‘genre’ and ‘season’ were converted to numerical type using One Hot Encoding Technique.



	visitors	ad_impressions	major_sports_event	views_trailer	views_content	genre_Comedy	genre_Drama	genre_Horror	genre_Romance	genre_Sci-Fi	genre_Thriller
0	1.67	1113.81	0	56.70	0.51	0	0	1	0	0	0
1	1.46	1498.41	1	52.69	0.32	0	0	0	0	0	1
2	1.47	1079.19	1	48.74	0.39	0	0	0	0	0	1
3	1.85	1342.77	1	49.81	0.44	0	0	0	0	1	0
4	1.46	1498.41	0	55.83	0.46	0	0	0	0	1	0

Fig. 4.4 Dummy Variables

5. MODEL BUILDING

- **Model Selection:**

An Ordinary Least Squares (OLS) regression model was chosen for this analysis due to its interpretability and suitability for examining the linear relationships between the target variable (views_content) and the predictor variables. This model allowed to quantify the impact of various factors, including major sports events, trailer views, and specific days and seasons, on the first-day viewership.

- **Training and Testing Split:**

The dataset was divided into 80% training and 20% testing sets, ensuring that the model was trained on the majority of the data while reserving a portion for evaluating its predictive performance on unseen data.

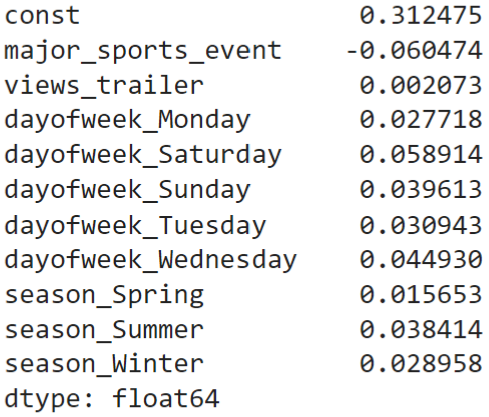
- **Removal of Non-significant Variables**

To enhance the model's interpretability and accuracy, non-significant variables were systematically removed. Variables with p-values above the commonly accepted threshold of 0.05 were excluded, as they did not contribute significantly to explaining the variance in viewership.

In addition to p-value analysis, multicollinearity among the predictor variables was assessed using the Variance Inflation Factor (VIF). Variables with high VIF values (above 5) were removed to ensure that multicollinearity did not distort the model's results. This step was crucial as multicollinearity can inflate the variance of coefficient estimates, making it difficult to assess the impact of individual variables accurately.

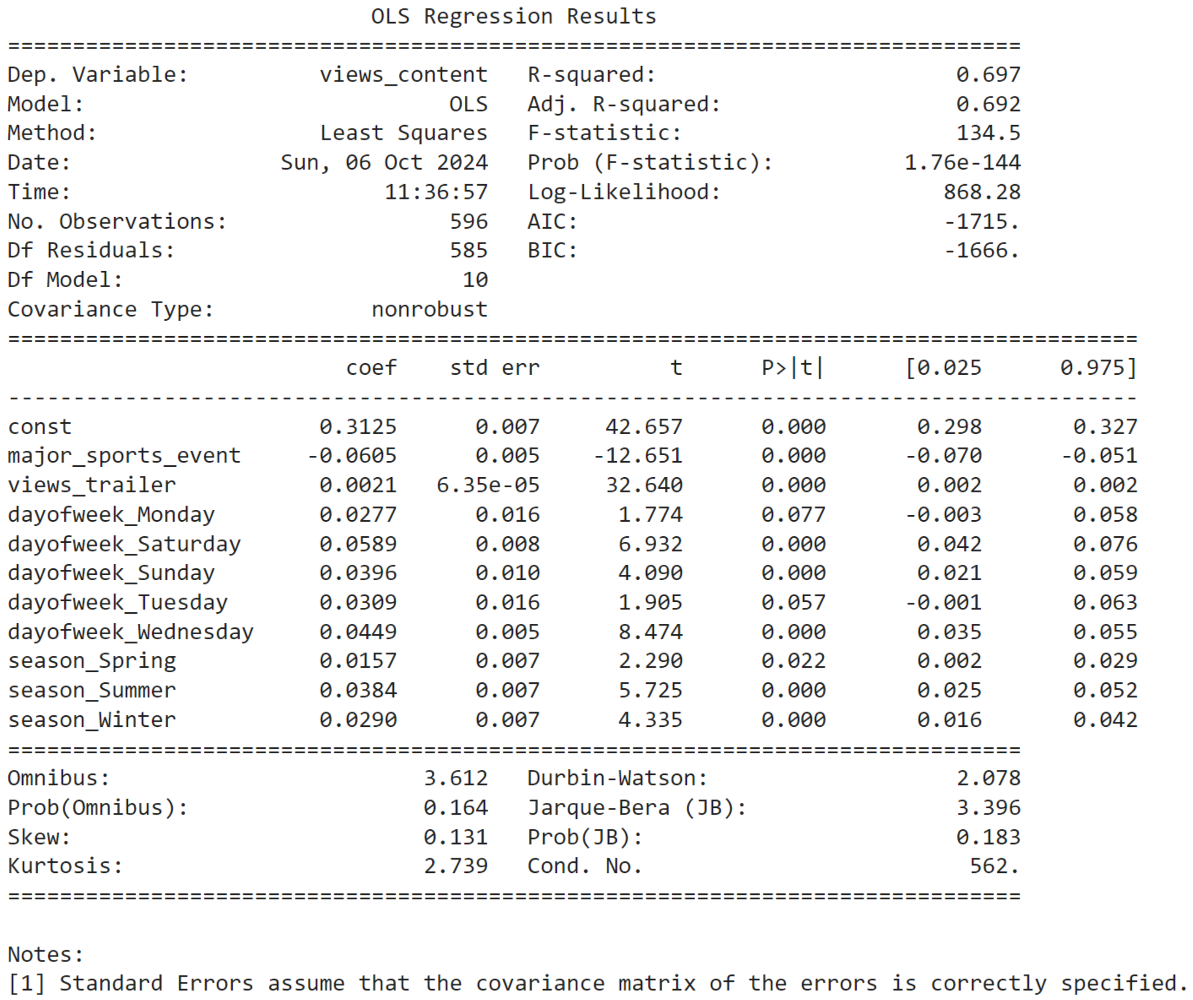
- **Evaluation Metrics and Residual Analysis**

The final model had an R-squared value of 0.697, indicating that nearly 70% of the variance in viewership could be explained by the included variables. Residual plots confirmed that the assumptions of normality and homoscedasticity were satisfied, further validated by the Durbin-Watson statistic, which was close to 2, suggesting minimal autocorrelation



```
const                  0.312475
major_sports_event    -0.060474
views_trailer          0.002073
dayofweek_Monday       0.027718
dayofweek_Saturday     0.058914
dayofweek_Sunday       0.039613
dayofweek_Tuesday      0.030943
dayofweek_Wednesday    0.044930
season_Spring          0.015653
season_Summer          0.038414
season_Winter          0.028958
dtype: float64
```

Fig. 5.1 Model Coefficients with Column names



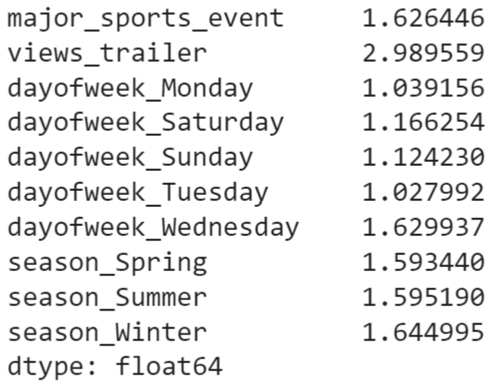
```
                            OLS Regression Results
==============================================================================
Dep. Variable:          views_content   R-squared:                       0.697
Model:                            OLS   Adj. R-squared:                  0.692
Method:                 Least Squares   F-statistic:                     134.5
Date:                Sun, 06 Oct 2024   Prob (F-statistic):          1.76e-144
Time:                        11:36:57   Log-Likelihood:                 868.28
No. Observations:                 596   AIC:                            -1715.
Df Residuals:                     585   BIC:                            -1666.
Df Model:                          10
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   0.3125      0.007     42.657      0.000       0.298       0.327
major_sports_event     -0.0605      0.005    -12.651      0.000      -0.070      -0.051
views_trailer           0.0021   6.35e-05     32.640      0.000       0.002       0.002
dayofweek_Monday        0.0277      0.016      1.774      0.077      -0.003       0.058
dayofweek_Saturday      0.0589      0.008      6.932      0.000       0.042       0.076
dayofweek_Sunday        0.0396      0.010      4.090      0.000       0.021       0.059
dayofweek_Tuesday       0.0309      0.016      1.905      0.057      -0.001       0.063
dayofweek_Wednesday     0.0449      0.005      8.474      0.000       0.035       0.055
season_Spring           0.0157      0.007      2.290      0.022       0.002       0.029
season_Summer           0.0384      0.007      5.725      0.000       0.025       0.052
season_Winter           0.0290      0.007      4.335      0.000       0.016       0.042
==============================================================================
Omnibus:                        3.612   Durbin-Watson:                   2.078
Prob(Omnibus):                  0.164   Jarque-Bera (JB):                3.396
Skew:                           0.131   Prob(JB):                        0.183
Kurtosis:                       2.739   Cond. No.                         562.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Fig. 5.2 OLS Regression Results

5.1 TESTING THE ASSUMPTIONS OF LINEAR REGRESSION MODEL

1. Checking for Multicollinearity



```
major_sports_event     1.626446
views_trailer          2.989559
dayofweek_Monday       1.039156
dayofweek_Saturday     1.166254
dayofweek_Sunday       1.124230
dayofweek_Tuesday      1.027992
dayofweek_Wednesday    1.629937
season_Spring          1.593440
season_Summer          1.595190
season_Winter          1.644995
dtype: float64
```

Fig. 5.1.1 Multicollinearity

- There is no multicollinearity in the model as it has been treated already

2. Checking for Linearity

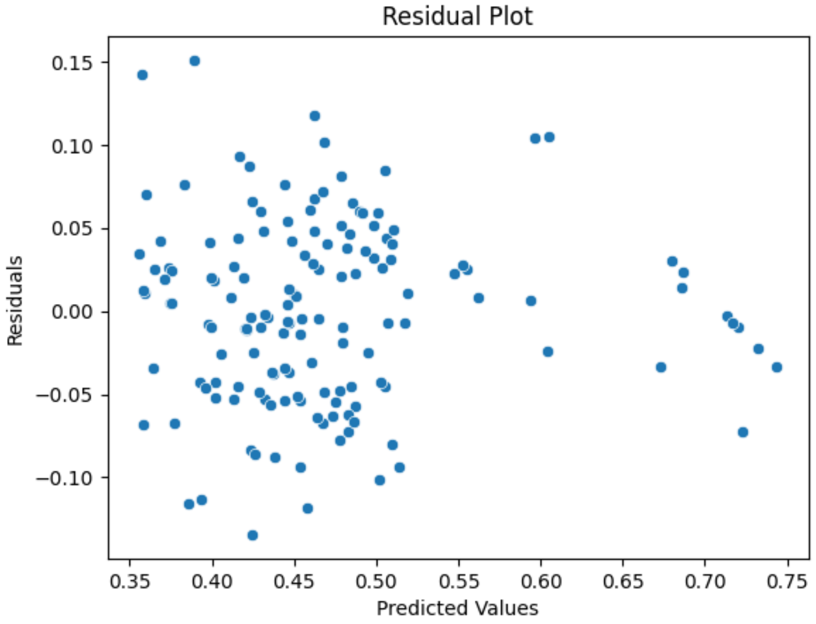


Fig. 5.1.2 Linearity and Independence

- There is no visible pattern in the Residuals plot indicating a Linear distribution.

3. Checking for Independence

- The Durbin-Watson statistic, which was close to 2, suggests minimal autocorrelation and thus, the assumption of independence is met.

4. Checking for Normality

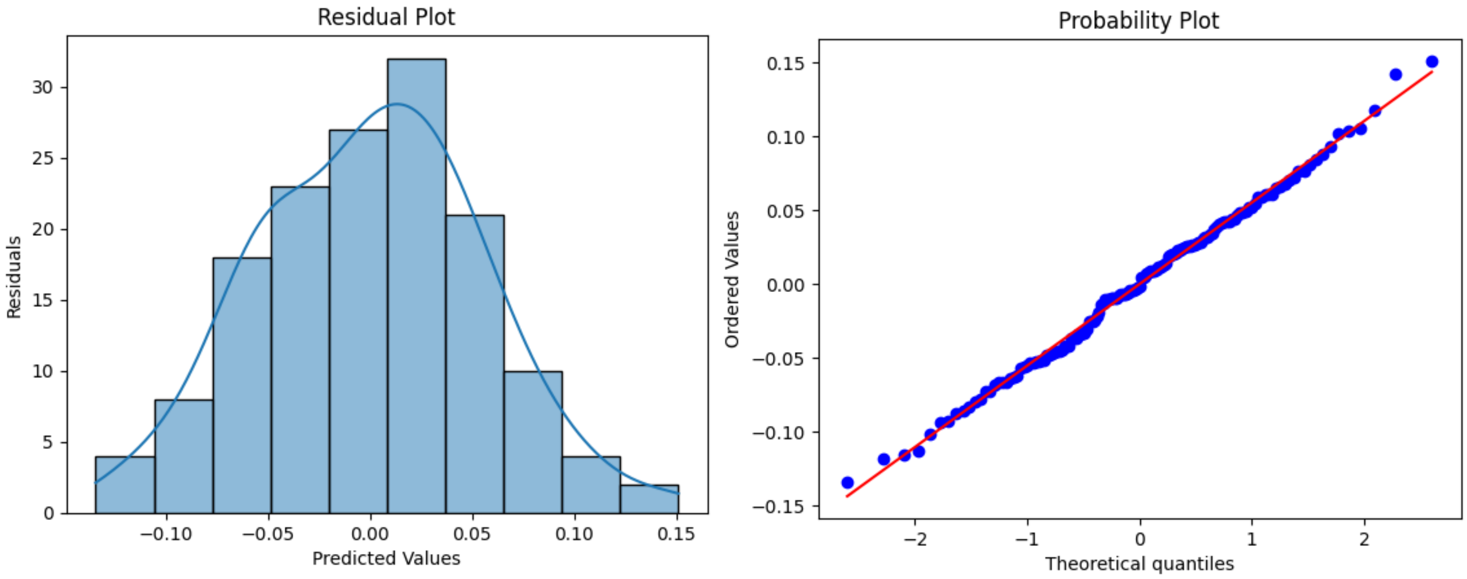
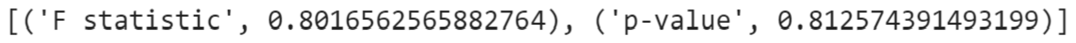


Fig. 5.1.3 Normality

- While not perfect, it can still be acceptable as a normal distribution

5. Checking for homoscedacity



```
[('F statistic', 0.8016562565882764), ('p-value', 0.812574391493199)]
```

- A **Goldfeld-Quandt test** was conducted to assess whether the variance of the residuals remains constant across observations. Since the resulting *p-value* was greater than 0.05, we fail to reject the null hypothesis of homoscedasticity. This indicates that there is no evidence of heteroscedasticity, and thus, we can conclude that the residuals are homoscedastic in this model.

5.2 MODEL PERFORMANCE EVALUATION

To assess the performance of the model, various metrics were examined, including R-squared, Adjusted R-squared, and the F-statistic.

- **R-squared**: The model achieved an R-squared of 0.697, indicating that approximately 69.7% of the variance in first-day content views is explained by the independent variables included in the model. This is a strong indication of the model's explanatory power.

- **Adjusted R-squared**: The Adjusted R-squared value of 0.692 accounts for the number of predictors in the model, confirming that the model retains a high level of explanatory power even after adjusting for the inclusion of multiple variables.

- **F-statistic**: With an F-statistic of 134.5 and a corresponding p-value less than 0.05, the overall model is statistically significant. This confirms that the predictors, as a whole, reliably predict the outcome variable.

- **Non-Significant Variables and Multicollinearity**:

During the model-building process, variables with high p-values (> 0.05) were iteratively removed to improve model precision and reduce complexity. This step ensured that only statistically significant predictors were included in the final model.

Additionally, multicollinearity was assessed using the Variance Inflation Factor (VIF), and variables exhibiting high VIF scores were excluded to ensure that multicollinearity did not bias the model estimates.

- **Residual Analysis**:

A residual analysis was conducted to check the model assumptions. The residuals displayed a random pattern, suggesting that the model adequately captures the relationship between the predictors and the dependent variable.

- **Homogeneity of Variance (Homoscedasticity)**:

A Goldfeld-Quandt test was performed to test for homoscedasticity. With a p-value greater than 0.05, we failed to reject the null hypothesis, indicating that the residuals are homoscedastic.

6 ACTIONABLE INSIGHTS AND BUSINESS RECOMMENDATIONS

ACTIONABLE INSIGHTS :

1. Content Release Timing:

- Most content is released on Fridays and Wednesdays, with the least on Mondays and Tuesdays. Weekends see a significant increase in viewership.

2. Seasonal Trends:

- Content releases peak during the fall and winter seasons, with higher median first-day viewership in winter and summer compared to spring and fall.

3. Genre Popularity and Strategy:

- Comedy is the most released genre, with Sci-Fi attracting the highest mean viewership. Romance performs best on Fridays, while Horror excels on Tuesdays.

4. Impact of Major Sports Events:

- The occurrence of major sports events significantly decreases the first-day content viewership.

5. Trailer Views and Content Promotion:

- There is a strong positive correlation between trailer views and first-day content views. Content with higher trailer views tends to have significantly higher first-day views.

6. Outliers in Viewership:

- Outliers are present on weekends, indicating some content gets unusually high first-day viewership on Saturdays and Sundays.

7. Content and Platform Visitors:

- There appears to be no direct correlation between overall platform visitors and first-day content views.

Inferences from OLS Regression Model:

- **Overall Model Performance:**

R-squared (0.697) indicates that around 69.7% of the variance in the dependent variable (first-day content views) can be explained by the predictors used in the model. This is a strong model fit.

F-statistic (134.5, p-value: 1.76e-144) confirms that the overall model is statistically significant, meaning the predictors collectively explain a significant portion of the variance in content views.

- **Impact of Trailer Views:**

Coefficient of views_trailer (0.0021, p < 0.001) suggests that for every additional trailer view, there is a 0.0021 increase in first-day content views, all else being equal. This is a very strong and positive relationship, making trailer promotion critical for boosting content views.

- **Impact of Major Sports Events:**

Negative coefficient of major_sports_event (-0.0605, p < 0.001) means that if a major sports event takes place, first-day content views decrease by approximately 6%. This is a substantial drop, suggesting that sports events have a large negative impact on viewership.

- **Day of Week Influence:**

Saturday (0.0589, p < 0.001) shows the strongest positive impact on first-day views, increasing them by about 5.89% compared to the reference day (probably Friday or some base day).

Sunday (0.0396, p < 0.001) also has a positive effect, increasing views by about 3.96%.

Wednesday (0.0449, p < 0.001) similarly contributes positively, increasing views by 4.49%.

Monday and Tuesday have weak positive impacts, but with higher p-values (0.077 and 0.057, respectively), making these days less significant statistically.

- **Seasonal Impact:**

Summer (0.0384, p < 0.001) and Winter (0.0290, p < 0.001) have positive impacts on viewership, with content in summer yielding a 3.84% increase and winter a 2.9% increase.

Spring (0.0157, p = 0.022) has the lowest positive effect but is still statistically significant.

RECOMMENDATIONS

Recommendations Based on Model Insights:

- **Focus on Trailer Marketing:**

Since trailer views significantly drive content viewership, invest heavily in promoting trailers across various platforms. Ensure trailers are engaging, properly marketed, and released well ahead of the content itself to maximize anticipation and viewership.

- **Avoid Major Sports Events:**

The significant negative impact of sports events on viewership (a 6% decrease) suggests avoiding content releases during such periods. Either avoid these dates altogether or use them for counter-programming with content that targets non-sports audiences.

- **Leverage Key Days for Content Release:**

The model clearly shows that Saturday, Sunday, and Wednesday are the best days to release content for maximizing first-day views. Scheduling high-profile content on these days can drive higher engagement. Monday and Tuesday are less impactful, so they should be reserved for lower-priority or niche content releases.

- **Maximize Seasonal Trends:**

Summer and Winter content releases are the most effective. These seasons likely coincide with holidays and increased leisure time, so focusing on blockbuster or highly anticipated content during these periods can significantly boost first-day views.

- **Explore Spring Releases:**

Though Spring has a weaker effect, it’s still statistically significant. Use spring releases for experimental or niche content that could benefit from less competition in these off-peak seasons.

- **Detailed Outlier Investigation:**

Investigating the outliers that occur on weekends, as noted in previous insights, may help in understanding what specific elements contribute to significantly higher-than-average views. These could include viral marketing, popular themes, or specific genres.